

AD-769 822

**AN APPROXIMATION FOR THE DISTRIBUTION
OF THE WILCOXON ONE-SAMPLE STATISTIC**

Winston K. Chow, et al

California University

Prepared for:

Office of Naval Research

1969

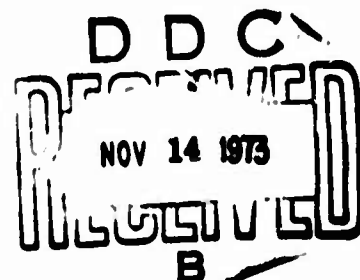
DISTRIBUTED BY:

NTIS

**National Technical Information Service
U. S. DEPARTMENT OF COMMERCE
5285 Port Royal Road, Springfield Va. 22151**

AN APPROXIMATION FOR THE DISTRIBUTION OF THE WILCOXON
ONE-SAMPLE STATISTIC

By Winston K. Chow¹ and J. L. Hodges, Jr.
University of California, Berkeley



1. Introduction

In 1945 Frank Wilcoxon suggested a statistic that may be used to test the location of a continuous symmetric population. Without loss of generality, we suppose that the null hypothesis puts the center of symmetry at zero, and that X_1, X_2, \dots, X_n are observations drawn from the population. Rank these observations in order of increasing absolute value, and attach to each rank the sign of the corresponding X_i . There are 2^n possible patterns of signs, and under the null hypothesis each pattern has the same probability $\frac{1}{2^n}$. Let W denote the sum of the ranks with negative sign. A small value of W constitutes evidence that the population center lies to the right of zero.

The Wilcoxon test, based on W , is attractive in several ways. The statistic itself is easy to compute. Its null distribution, which requires only the counting of the number, say $\#(w)$, of sign patterns giving $W = w$, provides exact significance probabilities without requiring any assumption of a parametric

¹This research was supported by the Office of Naval Research,
Contract NONR-NO0014-69A-0200-1038.

INTRODUCTION STATEMENT A

Approved for public release
Distribution Unlimited

40

form for the population (see Section 2). In normal samples, the Wilcoxon test has the high efficiency $3/\pi = .955$ in the limit as $n \rightarrow \infty$, relative to the t -test, and Klotz found in 1963 [5] that the efficiency is also high for normal samples of sizes $5 \leq n \leq 10$ provided α is not too small. For populations similar to the normal but with heavier tails, which is perhaps a typical situation in practical work, the large-sample efficiency of Wilcoxon relative to t can be arbitrarily greater than one.

Because of the importance of the Wilcoxon test, it is desirable to be able to find out something about the distribution of W in various circumstances. For example: Does the high efficiency in the normal case hold not only for $n \leq 10$ and $n \rightarrow \infty$, but also for moderate intermediate values? Does the asymptotic insensitivity of W to heavy tails hold also when n is small? How robust is the significance probability against moderate departures from symmetry?

For each of these and many similar questions, we need to find the distribution of W under the assumption that X_1, X_2, \dots, X_n are drawn from a population with distribution G which is not symmetric about zero. As reviewed below, this distributional question has not proved easy. We offer here a method of approximation which seems to be useful, at least in some cases, and then use it to throw some light on the specific questions asked above.

It is easy to see that what is said also applies, with appropriate modifications, to another important use for the Wilcoxon test-statistic--to test the absence of treatment effect in a matched-pairs design. In that case, X_i represents the observed difference in response between treated and control subject in the i th pair. The random assignment of treatments gives W its null distribution. If we imagine a population of pairs in which X has the distribution G , the questions of power and robustness may also be asked in this case. We shall however for simplicity couch our work in terms of the one-sample problem.

2. The null distribution

Under the null hypothesis, $P(W = w) = \#(w)/2^n$. The range of W is from 0 to $w_{hi} = \frac{1}{2}n(n+1)$, and the distribution is symmetric about $\frac{1}{2}n(n+1)$. The function $\#$, which is needed in our approximation, may be computed by means of recursion on n . Tables J and K of [2], provide all values of $\#(w)$ for $n \leq 12$; and for $w \leq n + 30$ where n ranges between 13 and 20, inclusive. For future reference, we show the values of $\#$ for $n = 10$ in Table 1. For $w \leq n$, values of $\#$ can also be read from the partition function q given in Table 24.5 of [7].

In case of large n , and $w > n$, the Edgeworth approximation leads to the approximation

$$\begin{aligned}
 (2.1) \quad \#(w) &= 2^n \cdot P(W = w) \\
 &= 2^n \cdot \varphi(z) \cdot \{1 - 3(u^2 - 6u + 3)/10 \cdot (2n+1) \\
 &\quad + 3(u^3 - 15u^2 + 45u - 15)/35n^2 \\
 &\quad + 9(u^4 - 28u^3 + 210u^2 - 420u + 105)/800n^2\} / \sigma
 \end{aligned}$$

where $\sigma^2 = n(n+1)(2n+1)/4$, $z = [w - \frac{1}{2}n(n+1)]/\sigma$ and $u = z^2$.

Formula (2.1) has relative error of order $1/n^3$, and Table 2 shows that it gives excellent results at the limits of the exact tables described above.

3. Distribution of W by numerical integration

Numerical integration gives a straightforward method of finding the distribution of W for a given G . Suppose the continuous G has a density g , which would be so in nearly all cases of interest. By independence, the joint density of the sample is $g(x_1)g(x_2)\cdots g(x_n)$. The n -dimensional space may be divided into 2^n regions corresponding to the 2^n patterns of signs for the absolute values. Evaluation of the integrals of this joint density over these 2^n regions will give probabilities, appropriate sums of which will give the distribution $P(W = w)$. An effective iterative scheme for evaluating these integrals was devised by Klotz [7], who in 1963 published results for normal samples with $5 \leq n \leq 10$. We are grateful to Professor Klotz for supplying us with his original results, only part of which have

been published. His values for $n = 10$ and $N(1,1)$ are shown in the third column of Table 4. Professor Klotz points out that the sum of his 1024 integrals is 1.0003, indicating that the fourth decimal place is not quite reliable. The Klotz iterative formulas were used by Arnold [1] in 1965 to obtain results for t-distributions with the same sample sizes.

While the integration approach is straightforward, it rapidly becomes very expensive as n increases. Not only does the number of regions increase exponentially with n , but even worse is the increase in dimensionality. (The difficulty of accurate integration in higher-dimensional space is discussed in detail by Milton [6] for the related two-sample problem.) If one wishes to survey a wide range of sample sizes and distributional shapes, integration does not appear to be a practical approach. Experience with other tests suggests that some sort of approximation is likely to be helpful as a supplement to the small-sample calculation. Fortunately, it is not difficult to find moments of W for use with approximations to its distribution.

4. Low-order moments of W

It is well known that W can be expressed as a sum of indicators, which leads to expressions for its moments in terms of certain probabilities. We will now record convenient formulas for the first three moments:

$$(4.1) \quad E(W) = \left[\frac{1}{2} q_1 \cdot (n-1) + p \right] \cdot n ,$$

$$(4.2) \quad \text{Var}(W) = ([m_{23} \cdot (n-2) + m_{22}](n-1) + m_{21})n,$$

$$(4.3) \quad \mu_3(W) = \langle ([m_{34} (n-3) + m_{33}](n-2) + m_{32}](n-1) + m_{31}) \rangle \cdot n,$$

where

$$(4.4) \quad m_{21} = p(1-p),$$

$$(4.5) \quad m_{22} = (p-q_1)^2 + 3q_1(1-q_1)/2,$$

$$(4.6) \quad m_{23} = q_2 - q_1^2,$$

$$(4.7) \quad m_{31} = p(1-p) \cdot (1-2p),$$

$$(4.8) \quad m_{32} = 6q_1^- (1-2p-q_1) - 6pq_1(1-q_1) \\ + 3p^2(1+3q_1) + \frac{1}{2}q_1 \cdot (1-q_1) \cdot (1-2q_1),$$

$$(4.9) \quad m_{33} = 6q_2^+ + 3q_2(2-3p-3q_1) + 6q_1^-(p-2q_1) \\ + q_1^2(12p+8q_1-3) + p^3,$$

$$(4.10) \quad m_{34} = q_3 + 3r - 9q_1q_2 + 5q_1^3,$$

and

$$(4.11) \quad p = P(X_1 < 0) = G(0)$$

$$(4.12) \quad q_k = \int_{-\infty}^{\infty} g(x) G^k(-x) dx$$

$$(4.13) \quad q_k^- = P(X_2, X_3, \dots, X_{k+1} < -X_1, X_1 < 0) \\ = \int_{-\infty}^0 G^k(-x) g(x) dx$$

and

$$(4.14) \quad q_k^+ = P(X_2, X_3, \dots, X_{k+1} < -X_1, X_1 > 0) \\ = \int_0^{\infty} G^k(-x) g(x) dx,$$

$$(4.15) \quad h(x) = g(x)G(-x),$$

$$(4.16) \quad H(x) = \int_{-\infty}^x h(u) du,$$

$$(4.17) \quad r = \int_{-\infty}^{\infty} h(x)H(-x) dx.$$

Notice that only univariate numerical integration is required by any of these formulas, and that all values of n are dealt with simultaneously. We give as Table 3 the values of the coefficients for the first three moments for the normal distribution $\mathcal{N}(\mu, 1)$ with unit variance and expectation μ , for some of the values of μ considered by Klotz.

5. The normal approximation

It follows from the work of Hoeffding [4] that, as $n \rightarrow \infty$, $\frac{W - E(W)}{S.D. (W)}$ will tend to the standard normal distribution. At first glance, one might hope that this fact, combining with small-sample integrations by the method of Klotz, would solve the problem. Unfortunately, it appears that the normal approximation is still very bad at the practicable limits of the integration approach. This is illustrated in the fourth column of Table 4, for a sample of $n = 10$ drawn from $\mathcal{N}(1, 1)$. The results of Klotz's integrations are compared with this normal approximation (with continuity correction) for $E(W) = 5.125\ 785$ and $\text{Var}(W) = 23.847\ 590$, which are the moments found by using Table 3 and formulas (4.1) and (4.2). The maximum error of the cumulative form of the normal approximation is -0.1004 .

The reason for the poor results is apparent if one glances at Figure 1, which shows the distribution of W as a histogram. The shape is quite unlike that of a normal curve, and indeed, without Klotz's work, the mere values of $E(W)$ and $S.D.(W)$ could have told us that this must be so. Since W is nonnegative, it cannot be nearly normal unless $E(W)$ is (say) at least 2.5 times as large as $S.D.(W)$. From (4.1) and (4.2), it can be shown that this will not occur until n reaches 67. The same general result is found for other population shapes than the normal. Therefore, this reveals that the normal tendency of W does not take effect until n is much larger than can practically be dealt with by integration. It is of course possible that some other "smooth" approximation may be found to give much better results. However, inspection of Figure 1, and similar figures for other cases, does not encourage one to hope for good small-sample results with approximations based on Edgeworth expansions or on the Pearson family of curves. We somehow need a method of approximation that deals with the irregularities of the null distribution of W .

6. The average probability method

Figure 1 shows that, for the case of a sample of 10 from $N(1,1)$, $P(W = w)$ tends to decrease as w increases, but in a rather irregular way. This irregularity can be explained by the irregularity of $\#(w)$ (Table 1). Thus, for each of the values $w = 0, 1$ and 2 , there is only one sign pattern, and these

Figure 1: Histogram of distribution
of W from $\mathcal{N}(1,1)$, $n = 10$

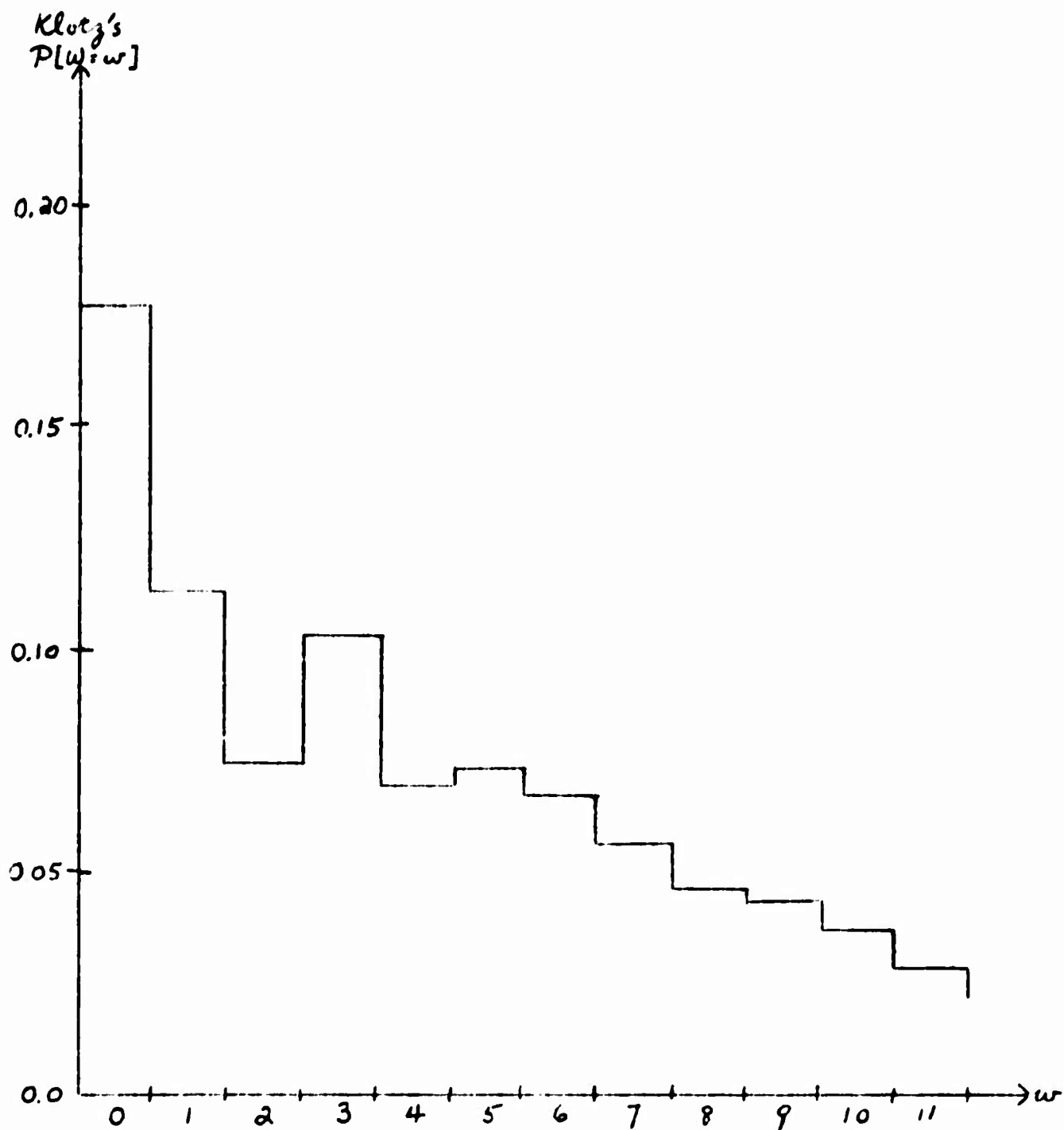


Figure 2: Average probabilities of W from $\mathcal{N}(1,1)$, $n = 10$

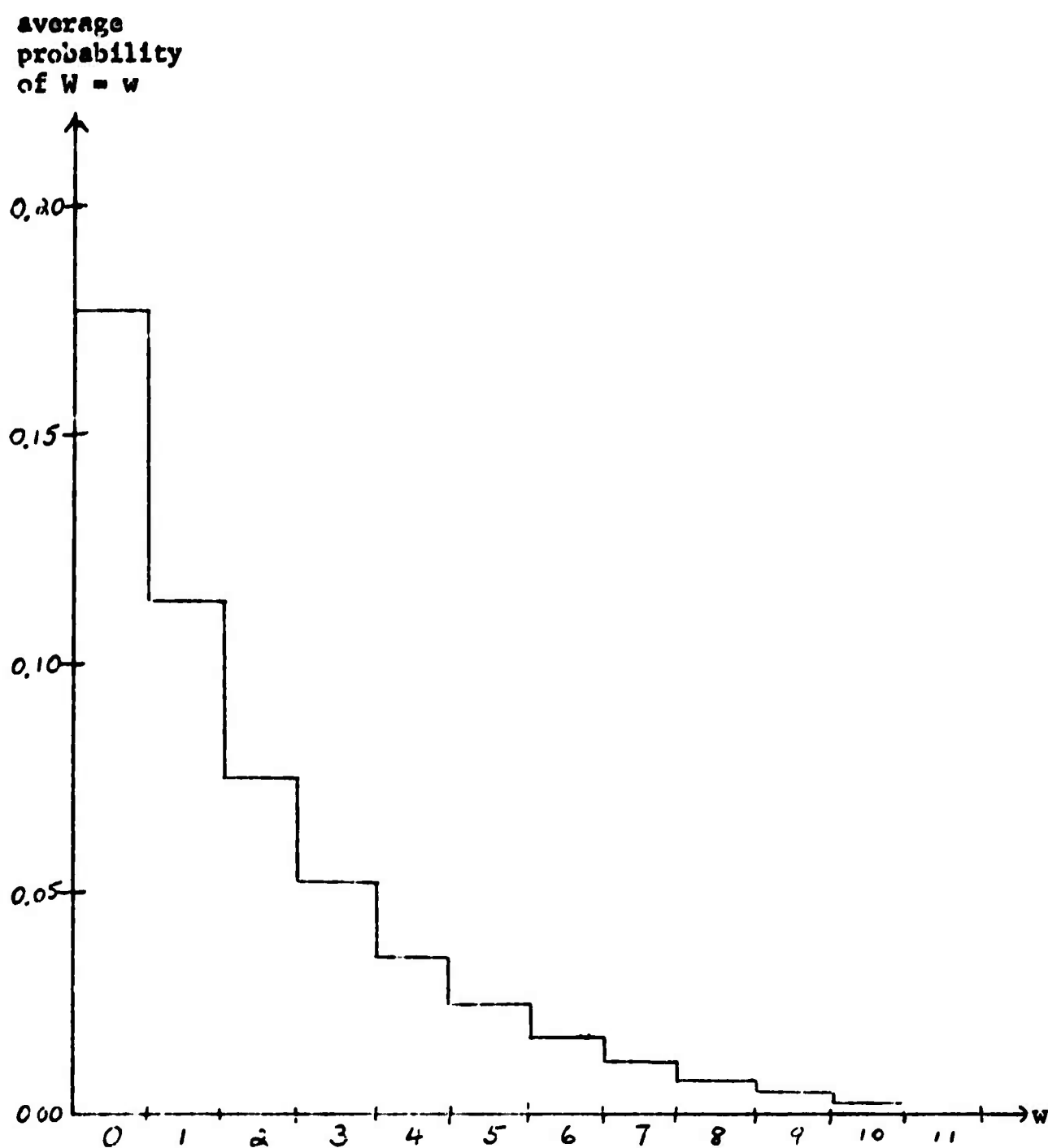
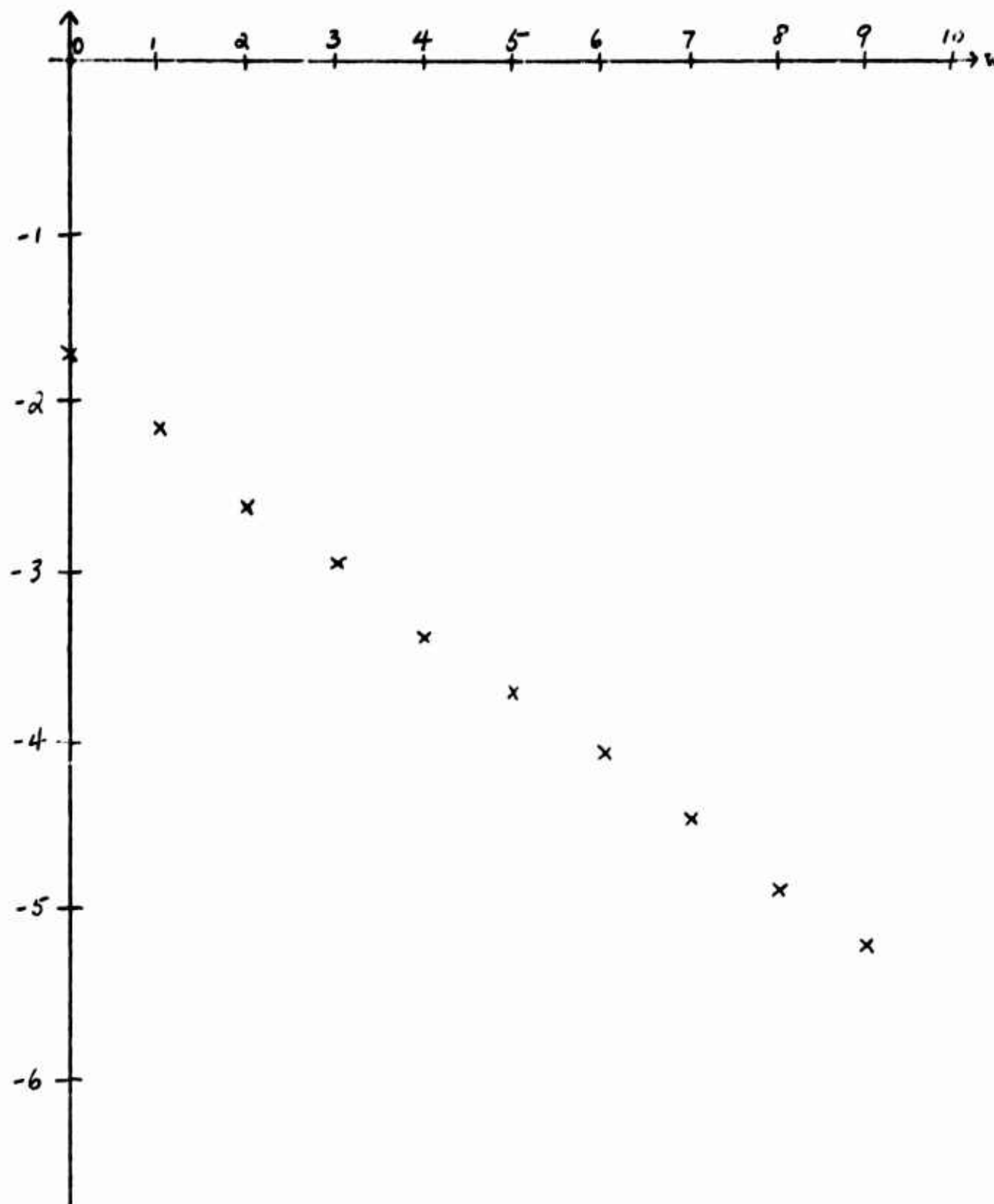


Figure 3: Log_e (average probability) from $\mathcal{M}(1,1)$, $n = 10$



probabilities decrease smoothly. At $w = 3$, there are however two sign patterns, which "explains" why $P(W = 3)$ exceeds $P(W = 2)$, rather than continuing to decline.

Let us consider $P(W = w)/\#(w)$, the average probability of the $\#(w)$ sign patterns corresponding to $W = w$. As Figure 2 shows, these average probabilities do decrease in a rather smooth way, at something like an exponential rate. This suggests that we write

$$(6.1) \quad L(w) = \log_e [P(W = w)/\#(w)],$$

and Figure 3 confirms the very smooth behavior, at least in this example, of L as a function of w .

We now have an idea for a method of approximation to the distribution of W . The values of L at the two extreme points are known: $W = 0$ if and only if all the n observations are positive, hence, we have $L(0) = n \log_e (1-p)$, since $\#(0) = 1$ and the probability of a positive observation is $1 - p$ by (4.11); similarly, $L(w_M) = n \log_e p$. By assuming that $L(w)$ behaves smoothly between these terminal values, we can interpolate some appropriate smooth function, say L' . If L' is close to L , then $P'(w) = \#(w) \cdot e^{L'(w)}$ will be close to $P(w)$. Various methods of fitting L' to L are possible. We have investigated the use of the known moments of W discussed in Section 4. The simplest functions with the required values at $w = 0$ and w_M are the

polynomials. Of course, some other smooth functions might be more appropriate in some cases. We shall however couch our discussion in terms of polynomial approximations which give satisfactory results for all cases we are going to consider in the next few sections.

We shall denote by L_k , the polynomial of degree k , fitted to L by the requirements (1) that $L_k(0) = L(0)$ and $L_k(w_M) = L(w_M)$, and (2) that the corresponding probabilities $P_k(w) = \#(w)e^{L_k(w)}$ have the same moments as W of orders $0, 1, \dots, k-2$. We thus impose $k + 1$ conditions, corresponding to the $k + 1$ coefficients of a polynomial of degree k .

First, let us consider the linear interpoland, $L_1(w) = L(0) + a_1 w$, where

$$(6.2) \quad a_1 = [L(w_M) - L(0)]/w_M = \frac{n}{w_M} \log \frac{p}{1-p}.$$

However, this interpoland will give probabilities $P_1(w)$ which in general do not add up to 1. A proper distribution requires use of the quadratic interpoland, which we shall write as

$$L_2(w) = L_1(w) + a_2 \cdot w \cdot (w_M - w).$$

Clearly, $P_2(w)$ is monotonely increasing in a_2 , so there exists a unique value of a_2 for which $\sum P_2(w) = 1$. Since the value of p determines L_1 , which in turn determines a_2 and hence L_2 . For any given n , P_2 is a one-parameter family of distributions, governed only by the parameter p , regardless of what the particular underlying distribution G is.

Table 5 shows, for $n = 10$ and for several values of p , the coefficients of L_2 , and the expectation $\sum w P_2(w)$ of the resulting distribution P_2 . In any given case, one may compare this expectation of the approximating P_2 with the true expectation $E(W)$, as given by (4.1), in order to help in judging the adequacy of the approximation P_2 .

As an illustration, let us consider a sample of 10 from $N(1,1)$. From a normal table one reads that $p = .1587$. Hence, from (6.2), $a_1 = -.303321$, also a_2 is uniquely determined, giving $a_2 = .001938$. The resulting probabilities P_2 have errors as shown in column 5 of Table 4. P_2 is substantially better than the normal approximation, having a maximum error only 12 percent as large. Of course, we know the error of P_2 only because the Klotz computations are available. However, even without those we could have considerable confidence in P_2 because of the readily available moment check. Interpolating the fifth column of Table 4 shows that the expectation of P_2 is 5.04, which is fairly close to the correct value $E(W) = 5.13$ obtained from (4.1).

In seeking a still better approximation, it is natural next to modify L_2 by adding a cubic term, to obtain an approximation with the correct expected value, say

$$L_3(w) = L_2(w) + a_3 \cdot w \cdot (w_M - w) \cdot (w - b_3) .$$

The proper values of a_3 and b_3 are determined by the requirements that $\Sigma P_3(w) = 1$ and $\Sigma w P_3(w) = E(W)$, which may be found numerically by inspection. (We have not proved the existence and uniqueness of a_3 and b_3 , but conjecture these properties will hold in all reasonable cases. At least in all of our trials the convergence to the fitted values proceeded smoothly and without difficulties.)

The search for a_3 and b_3 is aided by knowing where to start. Let us suppose that the P_2 approximation is a good one, so that only a small cubic correction is needed, and accordingly a_3 is small. Then, to a first approximation,

$$P_3(w) = P_2(w) + a_3 P_2(w) \cdot w(w_M - w) \cdot (w - b_3),$$

and this approximation, for each $W = w$, has error due to replacing $e^{\Delta(w)}$ by $1 + \Delta(w)$ for small $\Delta(w)$, where $\Delta(w) = a_3 \cdot w \cdot (w_M - w) \cdot (w - b_3)$. If we now impose the requirements that $\Sigma P_2(w) = \Sigma P_3(w) = 1$ and $\Sigma w P_3(w) = E(W)$, we find these approximation values for a_3 and b_3 :

$$b_3^* = S_1/S_0, \quad a_3^* = [E(W) - \Sigma w P_2]/(S_2 - b_3^* S_1)$$

where $S_k = \Sigma w^k P_2(w) \cdot w \cdot (w_M - w)$. Table 5 gives the values of b_3^* and $S_2 - b_3^* S_1$, which will indicate in any given case a good point for starting the search for a_3 and b_3 . We have written a program for the automatic conduct of the search. Once a_3 and b_3 are found, it is easy to compute L_3 and hence P_3 .

Table 4 shows the results of this method in our normal example. Here, $a_3^* = 0.000\ 013\ 3185$, $b_3^* = 9.0378$, and the searched $a_3 = 0.000\ 012\ 9195$, $b_3 = 9.1249$, and the agreement between P_3 and the Klotz values is better than that of P_2 .

The distribution P_3 will by definition have the correct values at $w = 0$ and at $w = w_M$, and the correct expectation. The quality of such an approximation may be judged by comparing EW^2 with $\sum w^2 P_3(w)$ or the corresponding variances. In this case, the variance of P_3 is 24.75, where the correct variance of W is 23.85. If the agreement is not as good as desired, one may add a quartic term to L_3 , getting

$$L_4(w) = L_3(w) + a_4 \cdot w \cdot (w_M - w) \cdot (w - b_4) \cdot (w - c_4),$$

where a_4 , b_4 , c_4 are determined by the requirements that P_4 have the correct moments of orders 0, 1, 2. In the cases we have examined, the quartic correction tends to be small, and to a good approximation with error discussed above, we obtain the trial coefficients given by:

$$b_4^* = \frac{1}{2S_1'^2 - 2S_2'S_0'} (S_2' \cdot S_1' - S_0' \cdot S_3' + [4S_2'^3 \cdot S_0' + 4S_3' \cdot S_1'^3 + (S_0' \cdot S_3')^3 - 3(S_2' \cdot S_1')^2 - 6S_0' S_1' S_2' S_3']^{\frac{1}{2}})$$

$$c_4^* = (S_2' - b_4^* \cdot S_1') / (S_1' - b_4^* \cdot S_0')$$

$$a_4^* = (EW^2 - \sum w^2 P_3(w)) / (S_2' - (b_4^* + c_4^*) S_1' + b_4^* c_4^* S_0')$$

where $S'_k = \sum w^k P_3(w) \cdot w \cdot (w_M - w)$. Unless one is lucky, the required moments corresponding to this set of trial coefficients need to be adjusted. One may make slight adjustments if these values do not give P_4 with moments sufficiently close to the true values as found by the formulas of Section 4.

The approximations P_2, P_3, P_4 are successively more accurate, as one might hope. The maximum absolute errors are 0.0120, 0.0083, and 0.0047 respectively. The root mean square errors for $1 \leq w \leq 20$ decrease also, being .0057, .0041, .0027.

In principle, one might now use $E(w^3)$ to add a quintic term, but if the method is working well for the case at hand, the agreement between P_3 and P_4 , and between EW^3 and $\sum w^3 P_4(w)$, should indicate that the P_4 approximation will serve. There is an intuitive reason for thinking that P_4 is a natural stopping point. Since both $\log P(w)$ and $\log \#(w)$ are nearly quadratic in w for large n , it follows from (6.1) that $L(w)$ will also be nearly quadratic in the interval near $E(W)$ containing most of the probabilities. For a polynomial to accord with this shape, and also have specified values at the "distant" points, $w = 0$ and w_M , it requires five degrees of freedom of a quartic. Thus, one may expect the approximation P_4 to agree with the asymptotic normal approximation in large samples, whereas P_3 could not be in general expected to do this.

Table 4 shows, for our normal example, the excellent agreement of P_4 with the results of the Klotz computations, with the maximum error of P_4 -cumulative is 0.0047. In addition, $\sum w^3 P_4(w) = 642.4183$, so that P_4 has standardized third moment 1.211, compared with the value 1.268 for W . This shows that the skewness of this approximating P_4 seems to be about right. While computations like those of Klotz will seldom be available for a check, this third-moment check can be made in general, as can the reassurance of finding P_4 close to P_3 .

We have also tried out the average-probability method on the computations of Arnold, for a sample of size $n = 10$ from a re-scaled t -distribution with four degrees of freedom, and for shift $\mu = 1.0$. By examining the probabilities published by Arnold, we concluded that for $w \geq 3$, they appear to have a smooth behavior. Hence, we applied our method to approximate distribution of W for $3 \leq w \leq 55$. For $w < 3$, one can always use the formulas stated in Section 7. The results are again good for P_4 , although not as good as in the normal example. The reason appears to be the exceptionally heavy tail of t_4 . The maximum difference between P_3 and P_4 is 0.0017, where the maximum error of the P_4 -cumulative is 0.008 from Arnold's results.

It is also necessary to point out situations when our approximation does not seem to work too well. One intuitive case we have at hand is $t_{\frac{1}{2}}$ with sample size $n = 10$ and shift $\mu = 1.0$.

After reviewing some of the probabilities published by Arnold, we realize that, with its extremely heavy tail, the assumption we made for our method, that L behaves smoothly between the two extremes, is not satisfied. As mentioned by Arnold, there is almost a complete breakdown in order for the case of $\frac{1}{2}$ degree of freedom in t-distribution. For instance, with $w = 9$, $\#(9) = 8$, there is one dominating factor corresponding to the sign pattern $(+, +, +, +, +, +, +, -, +)$ with probability 0.0498; when $w = 10$, $\#(10) = 10$, with dominating probability 0.1793, and when $w = 11$, $\#(11) = 11$, the dominating probability is only 0.0258. This irregularity of course lies on the fact that $t_{\frac{1}{2}}$ has an extraordinary heavy tail, and this leads us to believe that the average-probability method will not perform well in such cases.

7. An additional check

While computations by numerical integration of the entire distribution of W is seldom available and not easily done, it is not too difficult to get correct probabilities for small values of W . One can easily express these probabilities as univariate integrals:

$$(7.1) \quad P(W = 1) = n \int_{-\infty}^0 g(x) \cdot [1 - G(-x)]^{n-1} dx$$

$$(7.2) \quad P(W = 2) = n(n-1) \int_{-\infty}^0 g(x) \cdot [G(-x) - G(0)] \cdot [1 - G(-x)]^{n-2} dx$$

$$(7.3) \quad P(W = 3) = \frac{n(n-1)(n-2)}{2} \int_{-\infty}^0 g(x) \cdot [G(-x) - G(0)]^2 \cdot [1 - G(-x)]^{n-3} dx \\ + n(n-1) \int_{-\infty}^0 g(x) [G(0) - G(x)] \cdot [1 - G(-x)]^{n-2} dx.$$

We have found by numerical integration for the case of Table 4 that

$$P(W = 1) = .11365$$

$$P(W = 2) = .07501$$

$$P(W = 3) = .10360$$

which verify the results obtained by Klotz.

The probabilities for small w can be used to check on the adequacy of the L -approximations. Alternatively, they can be used to permit L to be interpolated between $w = 3$ and w_M rather than $w = 0$ and w_M . As in the t_4 -case of the preceding section, this is likely to be useful in irregular situations.

In a similar way, $P(W = w)$ for $4 \leq w \leq 8$ can be computed by bivariate integrals, and then L could be interpolated between $w = 8$ and $w = w_M$.

8. Normal samples of moderate size

The Pitman analysis shows that, with very large normal samples, the relative efficiency of the Wilcoxon test to the t -test is $3/\pi = 0.955$. Based on his computations for $5 \leq n \leq 10$, Klotz found that efficiencies lie in the range of $(.955, .986)$ for $.01 \leq \alpha \leq .10$. (The efficiencies were lower for very small α .) These facts have led to a widespread belief that the efficiency of the Wilcoxon test is high in all normal cases when the value of α is not too small. There has of course remained the possibility that the efficiency was less good for moderate values of n than for values at either extreme.

To throw some light on this question, we have applied the average-probability method to a sample of $n = 20$ from $(.75, 1)$. This computation requires $\#(w)$ for $n = 20$ recorded in Table 2. We found that $L(0) = -5.139\ 885$, $L(210) = -29.688\ 965$, $a_1 = -0.116\ 900\ 38$, $a_2 = -.000\ 135\ 9136$, $a_3 = -.000\ 000\ 2592$, $b_3 = 40.0236$, $a_4 = -.000\ 000\ 003\ 484$, $b_4 = 26.6508$, $c_4 = 62.8990$. Table 6 shows excerpts from the cumulative forms of the distribution P_3 and P_4 . The good agreement of these approximations, together with the satisfactory agreement of the standardized third moment for P_4 (0.644) and $W(0.664)$ lends confidence to the results.

As an additional check, we have evaluated (7.1), (7.2) and (7.3), with these results:

w	P(W = w)	
	by integration	approx. P_4
1	0.0050	.0051
2	0.0043	.0044
3	0.0074	.0076

From this P_4 -approximation, we find the following efficiencies for $n = 20$:

w	$\alpha =$	$\beta =$	β for t-test		Relative eff. of W to t
	$P_H[W \leq w]$	$P_A[W > w]$	n = 19	n = 20	
45	0.01198	0.20994	0.21888	0.19289	0.9672
52	0.02422	0.12760	0.13251	0.11415	0.9634
61	0.05270	0.06153	0.06335	0.05310	0.9588
66	0.07682	0.03938	0.04032	0.03328	0.9567
69	0.09467	0.02972	0.03049	0.02482	0.9568

9. The Wilcoxon distribution in a heavy-tailed case

Many statisticians consider that actual distributions found in practical work tend to resemble a normal, except that in some cases their tails are heavier than the very exiguous normal tails, corresponding perhaps to an occasional gross error. Such a departure from the normal shape can increase the population variance substantially, to the severe detriment of the large-sample performance of the t-test. On the other hand, this sort of departure from normality will have very little effect on the integral of the square of the population density, which governs the large-sample performance of the Wilcoxon test. Accordingly, the Pitman analysis shows that Wilcoxon can be substantially superior to t in such cases.

This is of course pertinent only when the sample is sufficiently large. If the sample is small, it would seem quite likely that none of the sample values will come from the tails, so that in effect one is sampling from the "normal" part of the

population. In that case the t-test should be superior. It is natural to pose a question: how large must the sample be before the heavy tails can exert their baleful effect on t relative to W ? An investigation of this important question has been hindered by the fact that it is difficult to calculate the power of either test. The average-probability method permits one to obtain a reasonably good approximation for the power of W . We present some results, partly in the hope that they may stimulate someone to think of a good way to do the same for the t-test, so that the comparison may be completed.

Let us take as our heavy-tailed distribution a Tukey model consisting of a blend of 97% from $\mathcal{N}(1,1)$ and 3% from a normal with the same expected value 1, but with standard deviation 4. For a sample of $n = 10$ from this distribution, we find that $L(0) = -1.814\ 432$, $L(55) = -17.961\ 627$, $a_1 = -0.293\ 585\ 36$, $a_2 = -.001\ 785\ 78$, $a_3 = 0.000\ 049\ 611\ 2$, $b_3 = 9.8276$, $a_4 = -.000\ 003\ 6214$, $b_4 = 6.2444$ and $c_4 = 18.6004$. Table 7 shows the cumulative forms of the distributions of P_3 and P_4 . Agreement is again good, also for the standardized third-moment check, with 1.122 for P_4 and 1.156 for W .

As an additional check, we used the integrals of Section 7:

w	$P(W = w)$	
	by integration	by P_4
1	.1043	.1058
2	.0690	.0700
3	.0954	.0942

All checks indicate that our P_4 -approximation is performing well in this heavy-tailed case.

With this blending of a Tukey model, the number of gross errors, G , is binomial ($n = 10$, $p = 0.03$) where $P(G = 0) = 0.737\ 424$. In other words, it may be quite likely, with probability $0.737\ 424$, that all the 10 sample values will be from pure $\mathcal{N}(1,1)$, for which case we have discussed in Section 6. One interesting point here is that, by subtracting this out, we are able to get the conditional distribution of W given $G > 0$. That is, knowing that some gross error occurred in this sample of 10,

$$P(W = w | G > 0) = \frac{1}{.262\ 576} [P_{\text{pure } \mathcal{N}(1,1)}(W = w) - .737\ 424 \cdot P_4(W = w)].$$

10. Robustness of the Wilcoxon test against asymmetry

The Wilcoxon one-sample test is intended to test the location of the center of symmetry of a symmetric population. Any actual population will be asymmetric, at least to some extent. It is accordingly important to know how robust the test is against asymmetry. That is, we need to find out how the actual significance probability compares with the nominal value given by the null distribution, in case the population is centered at zero but is moderately asymmetric. The average-probability method can throw some light on this important question.

Before this problem can be tackled, it is necessary to decide what is meant by the "center" of an asymmetric population.

The concept of center is naturally related to the statistical tool being used. If one works with the sample mean, then the population mean is the natural center, at least with large samples; similarly, the sample median calls for the population median; and so forth. What is the "Wilcoxon center" of a population?

Let us seek to define center in such a way as to promote the robustness of the significance probability. That is, we want the distribution of W , when the population is "centered" at 0, to resemble the null distribution in Section 2. This distribution is (except for very small n) nearly normal, so presumably W will continue to be something like normally distributed for mildly asymmetric populations. To keep the distribution of W nearly the same, we try to keep its location and scaling nearly the same.

In the symmetric case, if the null hypothesis is true, then $q_1 = \frac{1}{2}$ and $p = \frac{1}{2}$, so that

$$E(W) = [\frac{1}{2} q_1 (n-1) + p] \cdot n = \frac{1}{4} n(n+1).$$

One may seek to define center in general so that, under the null hypothesis, one will continue to have $q_1 = \frac{1}{2}$ and $p = \frac{1}{2}$, and thus W will continue to be centered at $\frac{1}{4} n(n+1)$, as in the symmetric case. Unfortunately, it turns out that this is in general impossible: if we locate the population so that $q_1 \approx \frac{1}{2}$, then p will differ at least slightly from $\frac{1}{2}$, and vice versa.

At least when n is large, the dominant term of $E(W)$ is the one involving q_1 . We are led to the idea of defining center so that, when the population is centered at 0, $q_1 = \frac{1}{2}$. Recalling that q_1 is the probability that $X_1 + X_2 < 0$ or equivalently that $\frac{1}{2}(X_1 + X_2) < 0$, we see that this event will have probability $\frac{1}{2}$ if the median of the distribution of $\frac{1}{2}(X_1 + X_2)$ is located at 0. We are led to define: The Wilcoxon center of a population is the median of the mean of two observations therefrom. We note that in the symmetric case, this definition yields the conventional center of symmetry.

In general, with this definition, p will differ slightly from $\frac{1}{2}$ for moderately asymmetric populations, and hence $E(W)$ will not quite coincide with $\frac{1}{4}n(n+1)$, though it will be so very near if n is large. However, the only way to force $E(W) = \frac{1}{4}n(n+1)$ would be to have the definition of the center of the population depend on the size of the sample, and that would be peculiar.

Let us illustrate these ideas. Consider the chi-square distribution with 10 degrees of freedom, depicted in Figure 4. This population has the sort of moderate skewness, with standardized third moment of 0.894 427, that might be encountered in practice in cases when the population was thought to be symmetric. If X_1 and X_2 are observations, independently, therefrom, then $X_1 + X_2$ has the chi-square distribution with 20 degrees of freedom

whose median is 19.337. Thus, the median of the distribution of $\frac{1}{2}(X_1 + X_2)$ is 9.6685, that is, $P(\frac{1}{2}(X_1 + X_2) < 9.6685) = \frac{1}{2}$, and the Wilcoxon center of χ_{10}^2 is at 9.6685. For comparison, the median is 9.6685, the mode is 8.00, and the expectation is 10.0, as illustrated in Figure 4. If we translate the

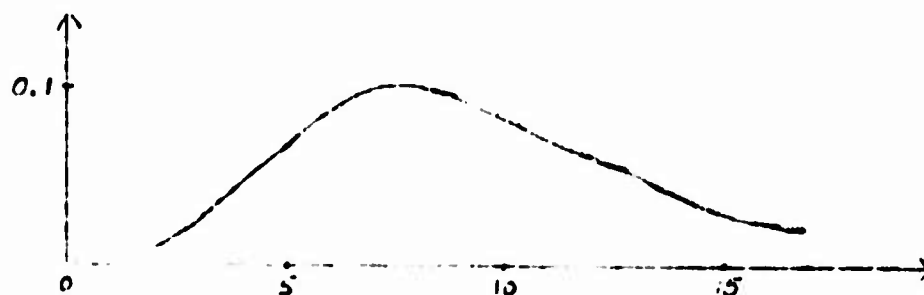


Figure 4

population to bring the Wilcoxon center to 0, then

$p = P(X_1 - 9.6685 < 0) = 0.53$. Accordingly, under the null hypothesis that the Wilcoxon center is at 0,

$$E(W) = \frac{1}{4}(n-1) \cdot n + 0.53 n = \frac{1}{4} n(n+1) \left(1 + \frac{4 \cdot (0.03)}{n+1}\right)$$

at $n = 10$, we get $E(W) = 27.8$ instead of the desired 27.5.

Since $S.D.(W)$ is approximately 9.8526, the discrepancy of 0.3 in the expected value of W is not especially important, and it would be even less so as n is increased above 10.

The magnitude of the departure of p from $\frac{1}{2}$ in typical cases can be assessed by examining the Edgeworth expansion:

$$\begin{aligned} p &= P[X_1 < \text{med}(\frac{1}{2}(X_1 + X_2))] = F_Z(z) \\ &= \Phi(z) - \eta_3 \varphi^{(2)}(z) + [\eta_4 \varphi^{(3)}(z) + \frac{1}{2} \eta_3^2 \varphi^{(5)}(z)] \\ &\quad - [\eta_5 \varphi^{(4)}(z) + \eta_3 \cdot \eta_4 \varphi^{(6)}(z) + \frac{1}{6} \eta_3^3 \varphi^{(8)}(z)] + \dots \end{aligned}$$

where

$$z = \frac{\text{med.} - EX_1}{S.D.(X_1)} .$$

In order to determine how far out should one carry for this expansion, one needs to examine the value determined in each batch ([]), which contains terms of the same order of convergence. If the value of a certain batch is small compared to the sum of the previous terms, one need not go any further. To illustrate this point, let us work out, for example, χ_{10}^2 with $n = 10$:

$$z = -0.074\ 126$$

where the sum of the first 4 terms sum up to 0.529 918. Adding in the next three terms will give $p = 0.530\ 418$. For our purpose, $p = 0.53$ will serve.

Let us now return to the robustness question. Suppose we have been given a specific asymmetric distribution, and have translated it to bring the Wilcoxon center to 0. This will make $q_1 = \frac{1}{2}$ and will yield a specific value of p . Accordingly, for a given sample size, $E(W)$ is determined. We may therefore apply the average-probability method to yield an approximating distribution P_3 . Therefore, we have in P_3 a one-parameter family of distributions determined by the parameter p . By the formulas of Section 4, we can calculate $\text{Var}(W)$ and $\mu_3(W)$. This permits the refinement of P_3 to P_4 , and a third-moment check on the adequacy of the P_4 approximation.

We have carried out this process for the χ^2_{10} example, for $n = 10$:

$$E(W) = (.249986 (n-1) + .529958) \cdot (n) = 27.8$$

$$\begin{aligned} \text{Var}(W) &= [(.084377 (n-2) + .375899)(n-1) + .249103] \cdot n \\ &= 97.0735 \end{aligned}$$

$$\begin{aligned} \mu_3(W) &= [[(.009280 (n-3) + .037334)(n-2) - .000113] \\ &\quad \cdot (n-1) - .014925] \cdot n \\ &= 73.49325 \end{aligned}$$

with standardized third moment 0.076842. We also found by the average-probability method,

$$\begin{aligned} L(0) &= -7.550\ 226, & L(55) &= -6.348\ 783 \\ a_1 &= 0.021\ 844\ 418, & a_2 &= -.000\ 007\ 4499, \\ a_3 &= -.000\ 037\ 473\ 2, & b_2 &= 28.32399 \\ a_4 &= .000\ 000\ 018\ 749, & b_4 &= 19.33199, & c_4 &= 36.44021. \end{aligned}$$

Table 8 shows excerpts of the P_4 -cumulative. At a first glance at some of the values shown, say, if you will reject if $W \leq 9$, then the significance probability shown by this approximation is 2.557%, where the nominal value given by the null distribution is 3.223%. This does not sound too appealing.

However, it is amazing how good the result will be, if one is doing a two-tailed test. For example, if one is willing to reject if $W \leq 9$ or $W \geq 46$, the nominal value will give 6.446%,

and the significance probability will add up to 6.609%. The reason that neither of the one-tailed test is too encouraging is due to the skewness based on this asymmetry distribution χ^2_{10} . But when one is doing a two-tailed test, the significance probabilities of both sides seem to balance out. Thus, one may conclude that the robustness property of Wilcoxon test against asymmetric distribution owns a much more sound evidence in the case of two-tailed test than either tail.

Table 1

#(w) for n = 10

w	#(w)	w	#(w)
0	1	14	17
1	1	15	20
2	1	16	22
3	2	17	24
4	2	18	27
5	3	19	29
6	4	20	31
7	5	21	33
8	6	22	35
9	8	23	36
10	10	24	38
11	11	25	39
12	13	26	39
13	15	27	40

Table 2. $\#(w)$ and approximation (2.1) for $n = 20$

w	$\#(w)$	$w \#(w)$ approx.	w	$\#(w)$	$w \#(w)$ approx.	w	$\#(w)$	$w \#(w)$ approx.
21	75	75.3	51	2131	2129.1	81	10538	10537.7
22	87	87.0	52	2300	2299.0	82	10864	10864.8
23	101	102.0	53	2479	2478.0	83	11186	11186.9
24	117	118.0	54	2668	2666.2	84	11504	11503.0
25	135	136.0	55	2865	2863.6	85	11812	11812.3
26	155	156.3	56	3071	3070.3	86	12113	12113.7
27	178	178.9	57	3288	3286.4	87	12407	12406.3
28	203	204.2	58	3512	3511.7	88	12689	12689.3
29	231	232.4	59	3746	3746.2	89	12961	12961.7
30	263	263.7	60	3991	3989.8	90	13224	13222.6
31	297	298.3	61	4242	4242.3	91	13471	13471.1
32	335	336.5	62	4503	4503.6	92	13706	13706.4
33	378	378.6	63	4774	4773.4	93	13929	13927.8
34	424	424.8	64	5051	5051.4	94	14134	14134.5
35	475	475.5	65	5337	5337.2	95	14326	14325.7
36	531	530.8	66	5631	5630.6	96	14502	14500.8
37	591	591.3	67	5930	5931.0	97	14659	14659.1
38	657	657.0	68	6237	6238.0	98	14800	14800.2
39	729	728.3	69	6551	6551.1	99	14925	14923.6
40	806	805.6	70	6869	6869.7	100	15029	15028.7
41	889	889.1	71	7192	7193.2	101	15115	15115.2
42	980	979.2	72	7521	7521.0	102	15184	15182.8
43	1076	1076.0	73	7851	7852.3	103	15231	15231.3
44	1180	1180.0	74	8185	8186.5	104	15260	15260.4
45	1293	1291.4	75	8523	8522.8	105	15272	15270.2
46	1411	1410.4	76	8859	8860.4			
47	1538	1537.4	77	9197	9198.5			
48	1674	1672.5	78	9536	9536.1			
49	1817	1816.1	79	9871	9872.5			
50	1969	1968.2	80	10206	10206.7			

Table 3. Coefficients of first three moments of W

coeff. μ	$(\mu, 1)$		$\mu = 0.25 \text{ (.25) } 1.50$			
	.25	.50	.75	1.00	1.25	1.50
p	.401 293	.291 160	.226 627	.158 655	.096 800	.066 807
q_1	.361 835	.234 235	.144 422	.078 650	.037 574	.016 947
m_{21}	.240 257	.206 386	.175 267	.133 484	.087 430	.062 344
m_{22}	.347 923	.277 200	.192 104	.115 097	.058 567	.027 476
m_{23}	.075 340	.056 585	.033 799	.016 881	.006 946	.002 403
m_{31}	.047 430	.086 203	.095 827	.091 128	.070 504	.054 014
m_{32}	.225 539	.345 618	.333 059	.243 020	.137 157	.070 001
m_{33}	.147 244	.201 906	.173 162	.103 590	.045 998	.017 046
m_{34}	.019 875	.030 189	.019 515	.009 968	.003 798	.001 058

Table 4. Comparison of Klotz's cumulative probabilities and normal approximation for $J(1,1)$, $n = 10$, together with the error shown for the approximating P_2 , P_3 and P_4

w	#(w)	Klotz $P(W \leq w)$	$10^4 \times \text{Error of}$			
			Normal	P_2	P_3	P_4
0	1	.1777	- 59	0	0	0
1	1	.2914	-.625	+ 45	+38	-27
2	1	.3664	- 710	+ 84	+70	-47
3	2	.4700	-1004	+105	+78	-40
4	2	.5397	- 907	+120	+83	-37
5	3	.6124	- 819	+114	+67	-18
6	4	.6804	- 696	+ 85	+31	-17
7	5	.7379	- 513	+ 65	+ 5	-37
8	6	.7847	- 295	+ 51	-11	-44
9	8	.8275	- 127	+ 39	-23	-45
10	10	.8638	+ 6	+ 34	-27	-36
11	11	.8914	+ 127	+ 30	-27	-26
12	13	.9146	+ 199	+ 21	-32	-21
13	15	.9328	+ 240	+ 18	-29	-11
14	17	.9476	+ 249	+ 12	-30	- 6
15	20	.9600	+ 232	+ 6	-31	- 4
16	22	.9693	+ 208	+ 4	-28	+ 1
17	24	.9764	+ 180	+ 3	-24	+ 6
18	27	.9822	+ 147	+ 1	-21	+ 7
19	29	.9866	+ 118	0	-18	+ 9
20	31	.9899	+ 93	0	-15	+ 9

Table 5. Given $n = 10$

p	a_2	$E(W)$	b_3^*	$s_2 - b_3^* \cdot s_1$
.02500	-.01634063	.31381	1.83707	22.15959
.05000	-.00904215	.79815	2.97823	169.80828
.07500	-.00579759	1.47716	4.28975	620.71346
.10000	-.00400116	2.34955	5.68875	1562.50147
.12500	-.00288598	3.39762	7.11726	3122.03702
.15000	-.00214043	4.59610	8.54549	5343.75854
.17500	-.00161469	5.91989	9.96180	8200.59779
.20000	-.00122953	7.34483	11.36261	11609.44424
.22500	-.00093940	8.84901	12.74807	15449.32008
.25000	-.00071599	10.41587	14.12031	19581.89779
.27500	-.00054129	12.03147	15.48157	23859.44368
.30000	-.00040333	13.68448	16.83394	28134.12860
.32500	-.00029388	15.36601	18.17928	32264.89875
.35000	-.00020715	17.06903	19.51912	36121.49424
.37500	-.00013905	18.78805	20.85475	39587.65504
.40000	-.00008661	20.51871	22.18717	42562.86995
.42500	-.00004772	22.25764	23.51725	44963.98652
.45000	-.00002075	24.00459	24.84574	46729.73432
.47500	-.00000517	25.75043	26.17310	47801.99147

Table 6. The cumulative probabilities of approximating

 P_3 and P_4 from $M(0.75, 1)$ with $n = 20$

w	$\#(w)$	$P_3(W \leq w)$	$P_4(W \leq w)$
0	1	.0059	.0059
1	1	.0109	.0109
2	1	.0153	.0153
3	2	.0230	.0229
4	2	.0296	.0295
5	3	.0382	.0381
6	4	.0481	.0480
7	5	.0589	.0587
8	6	.0701	.0699
9	8	.0831	.0828
10	10	.0972	.0968
11	12	.1119	.1113
12	15	.1278	.1272
13	18	.1444	.1437
14	22	.1620	.1611
15	27	.1807	.1798
16	32	.2000	.1990
17	38	.2200	.2188
18	46	.2409	.2397
19	54	.2623	.2610
20	64	.2843	.2829

Table 7. Excerpt of the cumulative probabilities of the approximating P_3 and P_4 from Tukey's model--a blend of 0.97 from $N(1,1)$ and 0.03 from $N(1,1)$; $n = 10$

w	$\#(w)$	$P_3(W \leq w)$	$P_4(W \leq w)$
0	1	.1629	.1629
1	1	.2707	.2687
2	1	.3426	.3387
3	2	.4396	.4330
4	2	.5055	.4973
5	3	.5734	.5642
6	4	.6360	.6266
7	5	.6906	.6818
8	6	.7367	.7291
9	8	.7803	.7744
10	10	.8192	.8154
26	39	.9952	.9977
27	40	.9964	.9984
28	40	.9973	.9989
29	39	.9980	.9993
30	39	.9985	.9995
31	38	.9989	.9997
32	36	.9992	.9998
33	35	.9994	.9999
34	33	.9996	.9999
35	31	.9997	1.0000

Table 8. P_4 -cumulative of Wilcoxon
distribution from χ^2_{10} with $n = 10$

W	N(W)	P_4 -cum	W	N(W)	P_4 -cum
0	1	.0005260	40	20	.8928811
1	1	.0010943	41	17	.9094992
2	1	.0017046	42	15	.9243857
3	2	.0030076	43	13	.9375258
4	2	.0043908	44	11	.9488886
5	3	.0065813	45	10	.9594839
6	4	.0096488	46	8	.9682115
7	5	.0136563	47	6	.9749787
8	6	.0186585	48	5	.9808340
9	8	.0255651	49	4	.9857195
10	10	.0344671	50	3	.9895591
11	11	.0445238	51	2	.9922548
12	13	.0566837	52	2	.9951082
13	15	.0709878	53	1	.9966265
14	17	.0874601	54	1	.9982513
15	20	.1070904	55	1	1.0000000

References

- [1] Arnold, H. J. (1965). Small sample power of the one sample Wilcoxon Test for non-normal shift alternatives. Ann. Math. Statist. 36, 1767-1778.
- [2] Hodges, J. L., Jr. and Lehmann, E. L. (1968). A compact table for power of the t-test. Ann. Math. Statist. 39, 1629-1637.
- [3] Hodges, J. L., Jr. and Lehmann, E. L. (1964). Basic Concepts of Probability and Statistics. Holden-Day, Inc.
- [4] Hoeffding, Wassily (1948). A class of statistics with asymptotically normal distribution. Ann. Math. Statist. 19, 293-325.
- [5] Klotz, J. H. (1963). Small sample power and efficiency for the one sample Wilcoxon and normal scores tests. Ann. Math. Statist. 34, 624-632.
- [6] Milton, R. C. (1967). Quadrature of high dimensional integrals: Applications to non-parameter statistics.
- [7] U. S. Department of Commerce, National Bureau of Standards (1964). Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. Appl. Math. Series 55.